МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ОДЕСЬКА ПОЛІТЕХНІКА»
Навчально-науковий інститут комп'ютерних систем
Кафедра інформаційних систем

Денис СІДОРОВ
(група АІ-224)

РОЗРАХУНКОВО-ГРАФІЧНА РОБОТА

З дисципліни «Методи та системи штучного інтелекту»

Моделювання задачі класифікації фільмів онлайн-кінотеатру за ознаками

Спеціальність:
F3 Комп'ютерні науки

Освітньо-професійна програма:
Комп'ютерні науки

Керівник:
Олена АРСІРІЙ, д.т.н., професор

Одеса – 2025

Мета розрахунково-графічної роботи. Закріплення знань та навичок по рішенню задач класифікації, регресії, прогнозування и кластеризації слабко структурованих вхідних даних за допомогою машинного навчання.

Завдання.

1. Зібрати, обробити вхідні данні для створення навчальної вибірки. Вимоги до вхідних даних наступні:
    - кількість аналізованих прикладів: не менше 60;
    - кількість аналізованих ознак: не менше 4-х.
2. Підготувати узгоджені данні для виконання машинного навчання.
3. Враховуючи практичні навички отримані при виконання лабораторних робіт обґрунтовано обрати необхідну модель для вирішення задачі і реалізувати її за допомогою обраної технології. Виконати навчання за допомогою обраної технології та підтвердити коректність отриманих результатів на основі тестування.
4. Для підтвердження адекватності створеної нейромережевої або будь-якої іншої моделі перевірити її дію на тестових прикладах.
5. Оформити пояснювальну записку.
6. Зробити висновки.
7. Підготувати презентацію результатів і публічно захистити РГР
8. Підготувати електронну версію РГР з робочим проектом створеної моделі вхідні данні навчальної та тестової вибірок, текст пояснювальної записки і презентацію.

Хід роботи.

Темою кваліфікаційної роботи є розробка системи для прогнозування бюджету фільму на основі непрямих даних. Створена система являє собою модель, навчену за допомогою Machine Learning, яка аналізує інформацію про фільм, відслідковує кореляції між даними та дає своє передбачення щодо бюджету.

За умовою роботи, дана таблиця фільмів (рис. 1). Необхідно, використовуючи бюджет (budget) як цільову змінну, провести класифікацію даних.

id	title	vote_average	vote_count	status	release_date	revenue	runtime	adult	backdrop_path	budget	homepage
27205	Inception	8.364	34495	Released	2010-07-15	825532764	148	False	/8ZTVqvKDQ8emSGUEMjsS4yHAwrp.jpg	160000000	https://www.warnerbros.com/movies/inc
157336	Interstellar	8.417	32571	Released	2014-11-05	701729206	169	False	/pbrkL804c8yAv3zBZR4QPEafpAR.jpg	165000000	http://www.interstellarmovie.net/
155	The Dark Knight	8.512	30619	Released	2008-07-16	1004558444	152	False	/nMKdUUepR0i5zn0y1T4CsSB5chy.jpg	185000000	https://www.warnerbros.com/movies/dar
19995	Avatar	7.573	29815	Released	2009-12-15	2923706026	162	False	/vL5LR6WdxWPjLPFRLe133jXWsh5.jpg	237000000	https://www.avatar.com/movies/avatar
24428	The Avengers	7.71	29166	Released	2012-04-25	1518815515	143	False	/9BBTo63ANSmhC4e6r62OJFuK2GL.jpg	220000000	https://www.marvel.com/movies/the-ave
293660	Deadpool	7.606	28894	Released	2016-02-09	783100000	108	False	/en971MEXui9diirXlogOrPKmsEn.jpg	58000000	https://www.20thcenturystudios.com/mo
299536	Avengers: Infinity War	8.255	27713	Released	2018-04-25	2052415039	149	False	/mDfJG3LC3Dqb67AZ52x3Z0jU0uB.jpg	300000000	https://www.marvel.com/movies/avenger
550	Fight Club	8.438	27238	Released	1999-10-15	100853753	139	False	/hZkgoQYus5vegHoetLkCJzb17zJ.jpg	63000000	http://www.foxmovies.com/movies/fight-
118340	Guardians of the Galaxy	7.906	26638	Released	2014-07-30	772776600	121	False	/uLtVbjvS1O7gXL8lUOwsFOH4man.jpg	170000000	http://marvel.com/guardians
680	Pulp Fiction	8.488	25893	Released	1994-09-10	213900000	154	False	/suaEOtk1N1sgg2MTM7oZd2cfVp3.jpg	8500000	https://www.miramax.com/movie/pulp-fi
13	Forrest Gump	8.477	25409	Released	1994-06-23	677387716	142	False	/qdIMHd4sEfJSckfVJfKQvisL02a.jpg	55000000	https://www.paramountmovies.com/mov
671	Harry Potter and the Philosopher's Stone	7.916	25379	Released	2001-11-16	976475550	152	False	/hziiv14OpD73u9gAak4XDDfBKa2.jpg	125000000	https://www.warnerbros.com/movies/har
1726	Iron Man	7.64	24874	Released	2008-04-30	585174222	126	False	/cyecB7godJ6kNHGONFjUyVN9OX5.jpg	140000000	https://www.marvel.com/movies/iron-ma
68718	Django Unchained	8.171	24672	Released	2012-12-25	425368238	165	False	/5Lbm0gpFDRAPIV1Cth6ln9iL1ou.jpg	100000000	http://www.unchainedmovie.com
278	The Shawshank Redemption	8.702	24649	Released	1994-09-23	28341469	142	False	/kXfqcdQKsToO0OUXHcrrNCHDBzO.jpg	25000000	
299534	Avengers: Endgame	8.263	23857	Released	2019-04-24	2800000000	181	False	/7RyHsO4yDXtBv1zUU3mTpHeQ0d5.jpg	356000000	https://www.marvel.com/movies/avenger
603	The Matrix	8.206	23815	Released	1999-03-30	463517383	136	False	/oMsxZEvz9a708d49b6UdZK1KAo5.jpg	63000000	http://www.warnerbros.com/matrix
597	Titanic	7.9	23637	Released	1997-11-18	2264162353	194	False	/rzdPqYx7Um4FUZeD8wpXqjAUcEm.jpg	200000000	https://www.paramountmovies.com/mov
475557	Joker	8.168	23425	Released	2019-10-01	1074458282	122	False	/hO7KbdvGOtDdeg0W4Y5nKEHeDDh.jpg	55000000	http://www.jokermovie.net/
120	The Lord of the Rings: The Fellowship of the Ring	8.402	23323	Released	2001-12-18	871368364	179	False	/x2RS3uTcsJJ9IfjNPcgDmukoEcQ.jpg	93000000	http://www.lordoftherings.net/
122	The Lord of the Rings: The Return of the King	8.474	22334	Released	2003-12-01	1118888979	201	False	/2u7zbn8EudG6kLlBzUYqP8RyFU4.jpg	94000000	http://www.lordoftherings.net
11324	Shutter Island	8.2	22318	Released	2010-02-14	294800000	138	False	/2nqsOT2AqPkTW81bWaLRtjgjqVM.jpg	80000000	http://www.shutterisland.com/
106646	The Wolf of Wall Street	8.035	22222	Released	2013-12-25	392000000	180	False	/63y4XSVTZ7mRzAzkqwi3o0ajDZZ.jpg	100000000	http://www.thewolfofwallstreet.com/
99861	Avengers: Age of Ultron	7.276	21754	Released	2015-04-22	1405403694	141	False	/6YwkGolwdOMNpbTOmLjoehlVWs5.jpg	365000000	http://marvel.com/movies/movie/193/ave
271110	Captain America: Civil War	7.4	21541	Released	2016-04-27	1155046416	147	False	/wdwcOBMkt3zmPQuEMxB3FUtMio2.jpg	250000000	https://www.marvel.com/movies/captain-

Рисунок 1 – Фрагмент таблиці

Для початку необхідно сформувати вхідні дані – вони будуть обрані та нормалізовані спеціально для виконання РГР. Цільову змінну budget розділимо на 6 класів:

– F tier (бюджет < 5.000.000);

– D tier (бюджет від 5.000.000 до 15.000.000);

– C tier (бюджет від 15.000.000 до 25.000.000);

– B tier (бюджет від 25.000.000 до 50.000.000);

– A tier (бюджет від 50.000.000 до 100.000.000);

– S tier (бюджет > 100.000.000);

В якості ознак для класифікації використаємо наступні змінні:

– жанри (категоріальна змінна, яка в процесі роботи буде розбита і перетворена на числову);

– країни виробництва (категоріальна змінна, яка в процесі роботи буде розбита і перетворена на числову);

– компанії-виробники (категоріальна змінна, яка в процесі роботи буде розбита і перетворена на числову);

– тривалість фільмів (невід'ємна числова змінна).

В якості класифікаторів було обрано дерево рішень, нейронну мережу, логістичну регресію та метод k-середніх. Для вирішення задачі скористаємося засобами пакету Orange. Послідовність віджетів у середовищі Orange для вирішення задачі на даних фільмів наведена на рисунку 2.

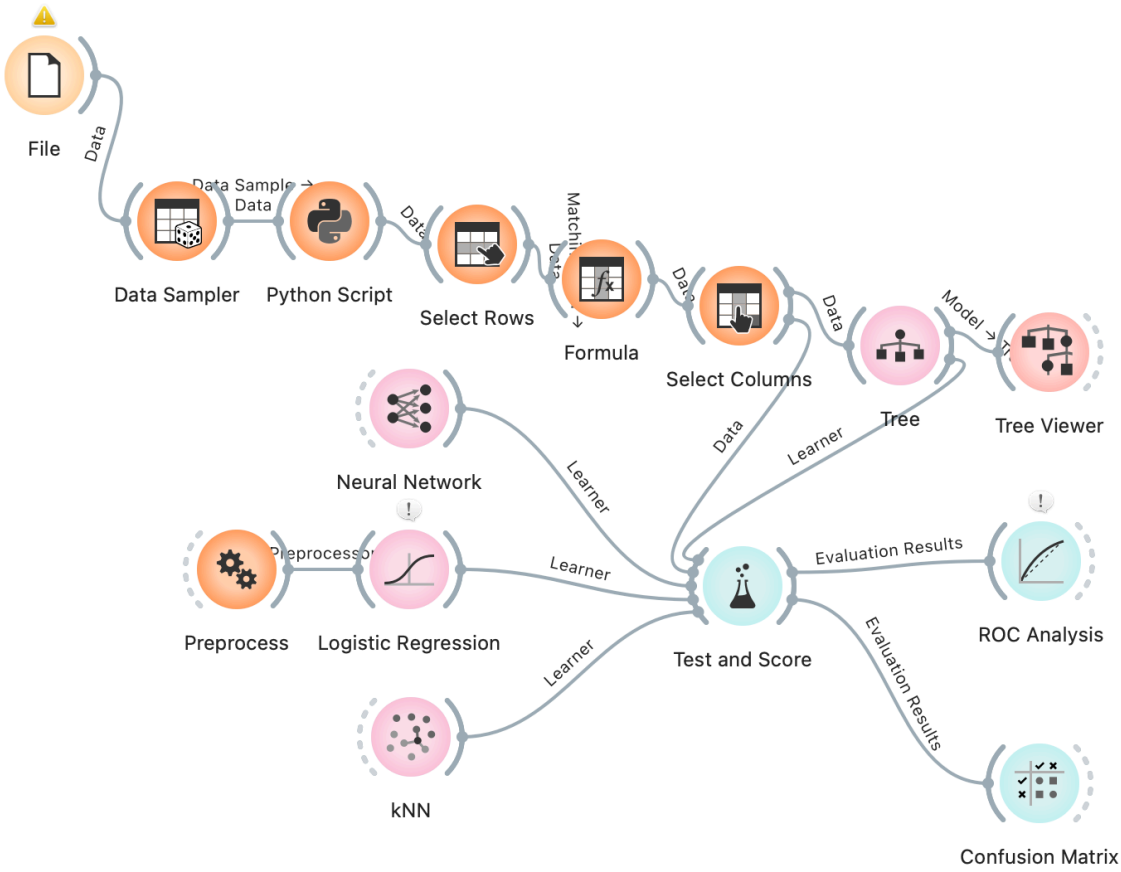


Рисунок 2 – Скріншот послідовності віджетів робочого процесу

Спочатку завантажимо датасет фільмів через віджет File, вкажемо змінні feature і targer та за допомогою Data Sampler випадково виберемо 10.000 прикладів.

В нашому датасеті категоріальні змінні genres, production_countries та production_companies можуть містити декілька значень через кому. Модель не

зможе правильно обробити такі дані і це призведе до суттєвого зниження якості, тому нам необхідно розділити їх та перетворити в числові змінні формату 0 (false) або 1 (true). Для цього напишемо окрему функцію і вбудуємо її в нашу модель через віджет Python Script (лістинг 1).

Лістинг 1 – Код для перетворення категоріальних змінних

```
import pandas as pd
from Orange.data import Table, Domain, ContinuousVariable
# 1. Зчитуємо дані з Orange
df = pd.DataFrame(in_data.X, columns=[v.name for v in
in_data.domain.attributes])
# Додаємо target (якщо є)
if in_data.domain.class_var:
   df[in_data.domain.class_var.name] = in_data.Y
# Додаємо meta-стовпці
for m in in_data.domain.metas:
   df[m.name] = in_data.get_column(m)
# ---- 2. Колонки для розбиття ----
columns_to_split = {
   "genres": "genre_",
   "production_countries": "country_",
   "production_companies": "company_",
}
for col, prefix in columns_to_split.items():
   if col in df.columns:
       df[col] = df[col].fillna("")
       df[col] = df[col].apply(lambda x: [s.strip() for s in
str(x).split(",") if s.strip()])
       # Створюємо one-hot з префіксом
       dummies = (
           df[col].explode()
                  .str.get_dummies()
                  .groupby(level=0)
                  .sum()
       )
       # Додаємо префікс до назв колонок
       dummies = dummies.add_prefix(prefix)
```

```
        # Замінюємо колонку на one-hot
        df = pd.concat([df.drop(columns=[col]), dummies], axis=1)
# ---- 3. Формуємо домен (тільки числові ознаки) ----
new_vars = []
for col in df.columns:
    new_vars.append(ContinuousVariable(col))
domain = Domain(new_vars)
# ---- 4. Повертаємо в Orange ----
out_data = Table.from_numpy(domain, df.to_numpy())
```

Також за допомогою віджету Select Rows позбавимося нульових (порожніх) змінних у стовпцях budget та runtime, щоб вони не вносили шум в модель.

Далі використаємо віджет Formula для перетворення числової цільової змінної budget в категоріальну відповідно до меж, сформованих на початку роботи (рис. 3) і внесемо її в модель замість старої числової змінної через віджет Select Columns.

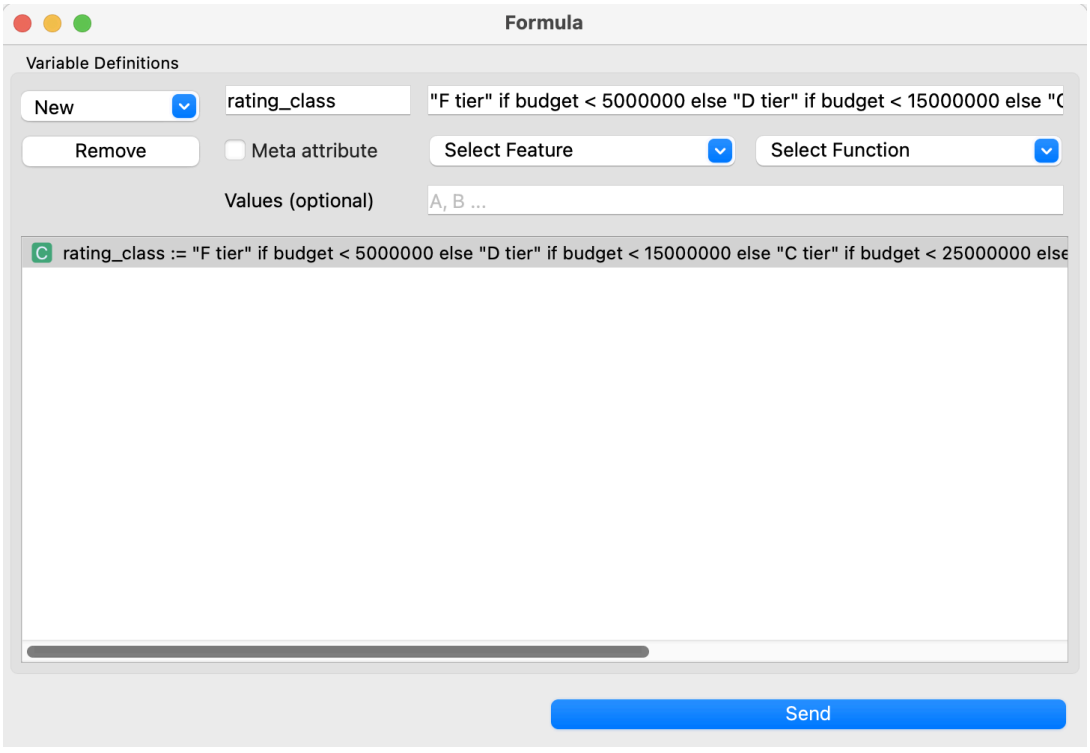
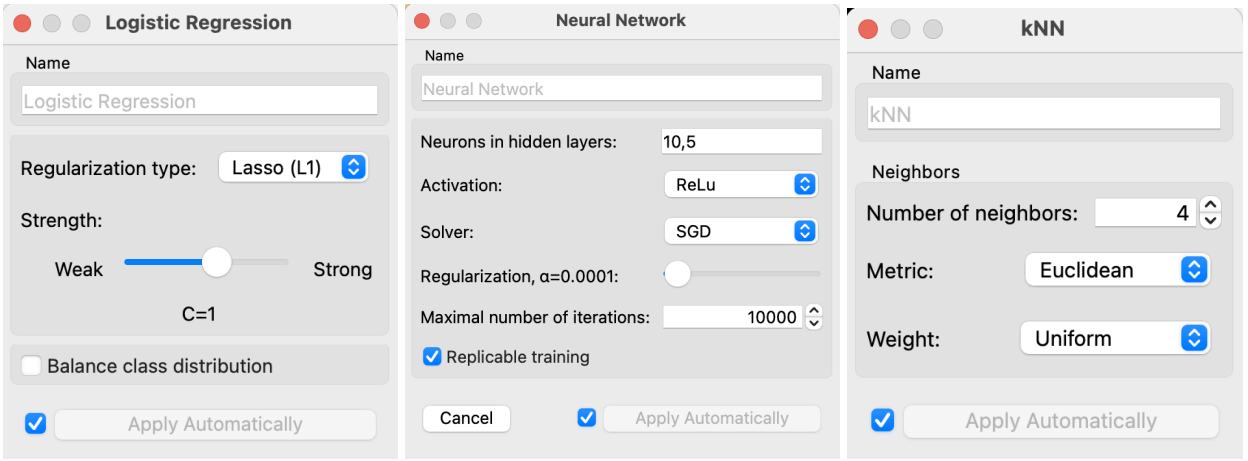


Рисунок 3 – Скріншот віджету Formula

З групи Model виберемо віджети Logistic Regression, Neural Network та kNN і налаштуємо відповідні параметри (рис. 4 а, б і в).



а б в

Рисунок 4 – Скріншоти параметрів віджетів Logistic Regression (а), Neural Network (б) та kNN (в)

Після проведення експериментів, у віджеті логістичної регресії було встановлено регуляризацію Lasso (L1), яка обнуляє менш важливі ваги, виконуючи відбір ознак. Параметр C=1 задає середню силу регуляризації. Для нейронної мережі було задано два приховані шари з кількістю нейронів 10 і 5 відповідно. ReLU використовується як функція активації, яка добре працює з нелінійностями та пришвидшує навчання. Solver SGD означає стохастичний градієнтний спуск: він повільніший, але більш контрольований і добре реплікується. У kNN встановлена кількість сусідів k=4, що робить модель досить чутливою до локальних шумів, але не перенавчає її. Для вимірювання близькості використовується евклідова відстань, а параметр Weight=Uniform означає, що всі сусіди голосують однаково, незалежно від того, наскільки вони близькі до точки. Саме ці значення дали найточніші результати.

До віджету Logistic Regression підключимо віджет Preprocess та встановимо нормалізацію значень ознак до інтервалу від 0 до 1 (рис. 5).

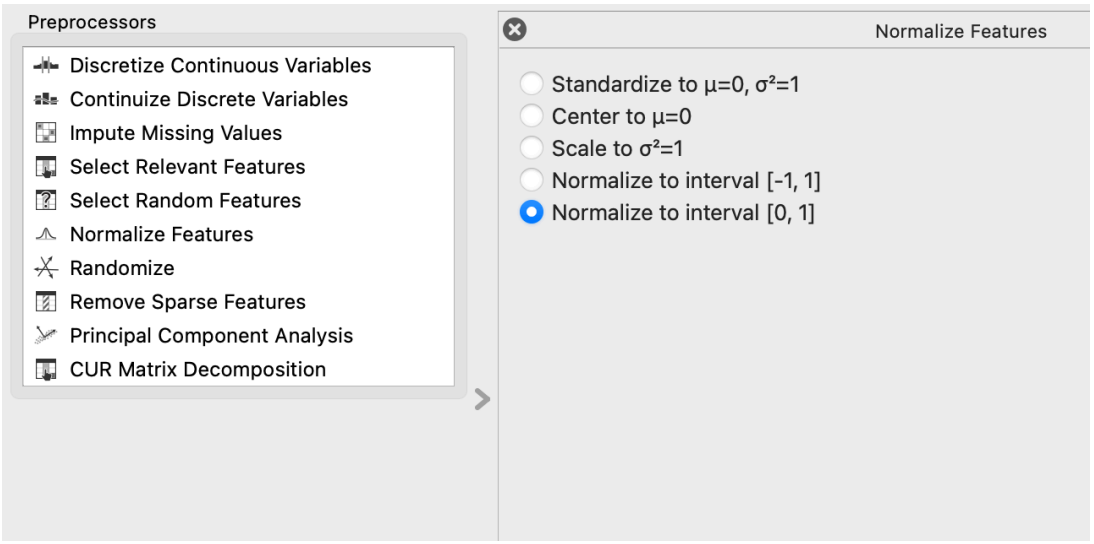


Рисунок 5 – Скріншот віджету Preprocess

Також додамо віджет Tree і налаштуємо його параметри (рис. 6).

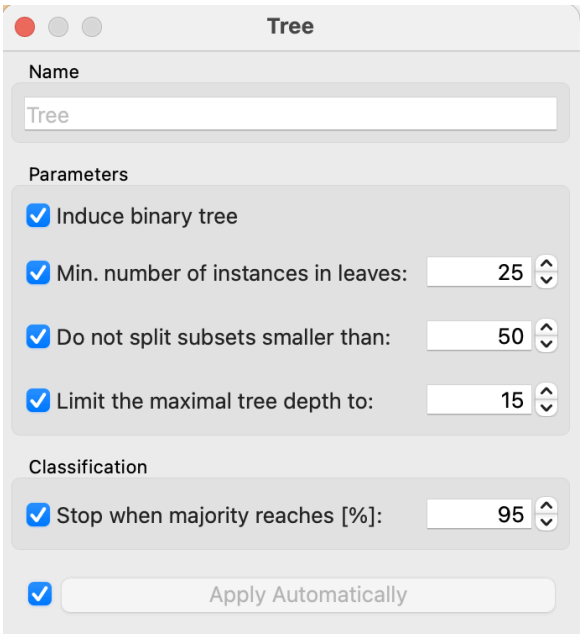


Рисунок 6 – Скріншот параметрів віджету Tree

Параметр Min. number of instances in leaves=25 не дає дереву створювати надто дрібні листки, захищаючи його від перенавчання.

Обмеження Do not split subsets smaller than 50 забороняє робити нові розгалуження, якщо підмножина даних уже мала, що також зменшує шумові спліти. Додатково обмежено максимальну глибину дерева – 15, щоб модель не зростала безконтрольно у глибину.

У віджеті Tree Viewer бачимо результат роботи дерева рішень (рис. 7).

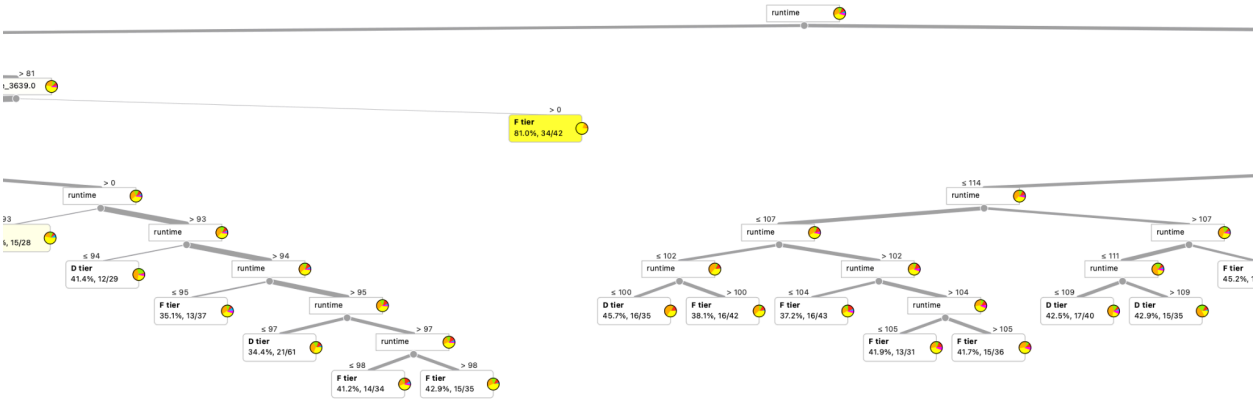


Рисунок 7 – Скріншот результатів з віджету Tree Viewer

Тепер візьмемо один об’єкт з вибірки і, згідно зі структурою дерева рішень, продивимось просування даного об’єкта від кореневого вузла до листа, в якому об’єкт буде класифікований, дотримуючись умов переходу, що визначають по якому із ребер йти. Візьмемо об’єкт з рядка 48 нашої вибірки даних (табл. 1)

Таблиця 1 – дані вибраного об’єкту

title	runtime	budget	genres	production_companies	production_countries
Harry Potter and the Goblet of Fire	157	150.000.000	Adventure, Fantasy	Warner Bros. Pictures, Heyday Films, Patalex IV Productions Limited	United Kingdom, United States of America

Першим беремо атрибут runtime. Він = 157, це > 99, тому переходимо по правому ребру. Наступний атрибут – country_2683.0 (США): 0 (false) чи 1 (true). Він true, тому йдемо по правому ребру вниз. Наступний знову runtime. 157 > 126, тому йдемо далі по правому ребру. Знову перевіряємо runtime – 157 > 140 – переходимо по правому ребру і потрапляємо до результату: S tier (бюджет > 100 мільйонів) з ймовірністю в 27,3%. Бюджет цього фільму справді становить більше 100 мільйонів, отже, класифікацію виконано правильно.

Виведемо результати роботи всіх класифікаторів за допомогою віджету Test and Score (рис. 8).

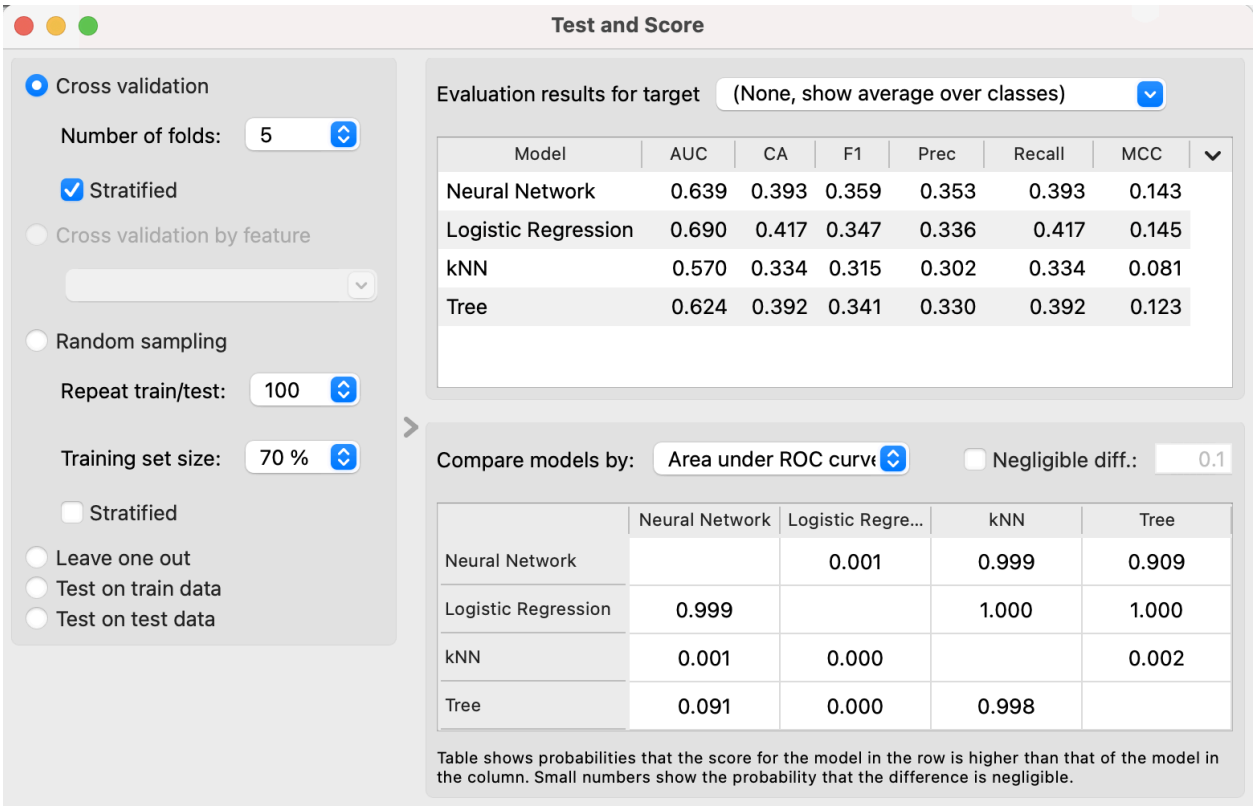


Рисунок 8 – Скріншот віджету Test and Score

Згідно з результатами модуля Test & Score, Logistic Regression є найстабільнішим і найефективнішим класифікатором у цій задачі: вона демонструє найвищий AUC (0,690 або майже 70%), а також найкращі показники CA, Recall та MCC, що говорить про її здатність краще за інші

методи розрізняти класи в умовах складних залежностей. Neural Network показує трохи нижчі результати (AUC 0,639 або майже 64%), але тримається на другому місці, тоді як дерево класів має менш стійкі метрики (AUC 0,624 або 62,5%), що властиво деревам при роботі з шумними або розрідженими ознаками. Найгірше поводиться kNN, який демонструє найнижчий AUC (0,570 або 57%) та загальні результати: модель слабко узагальнює дані, не може відслідкувати залежності між ними та в цілому видає результат не набагато кращий за випадковий. Матриця порівняння моделей підтверджує отримані результати: у 99,9% повторів Logistic Regression перевершує нейромережу і завжди перевершує kNN та Tree, тоді як kNN програє всім іншим моделям.

Перейдемо до віджета ROC Analysis. Виведемо графіки ROC-кривих по одному графіку на клас (рис. 9-14).

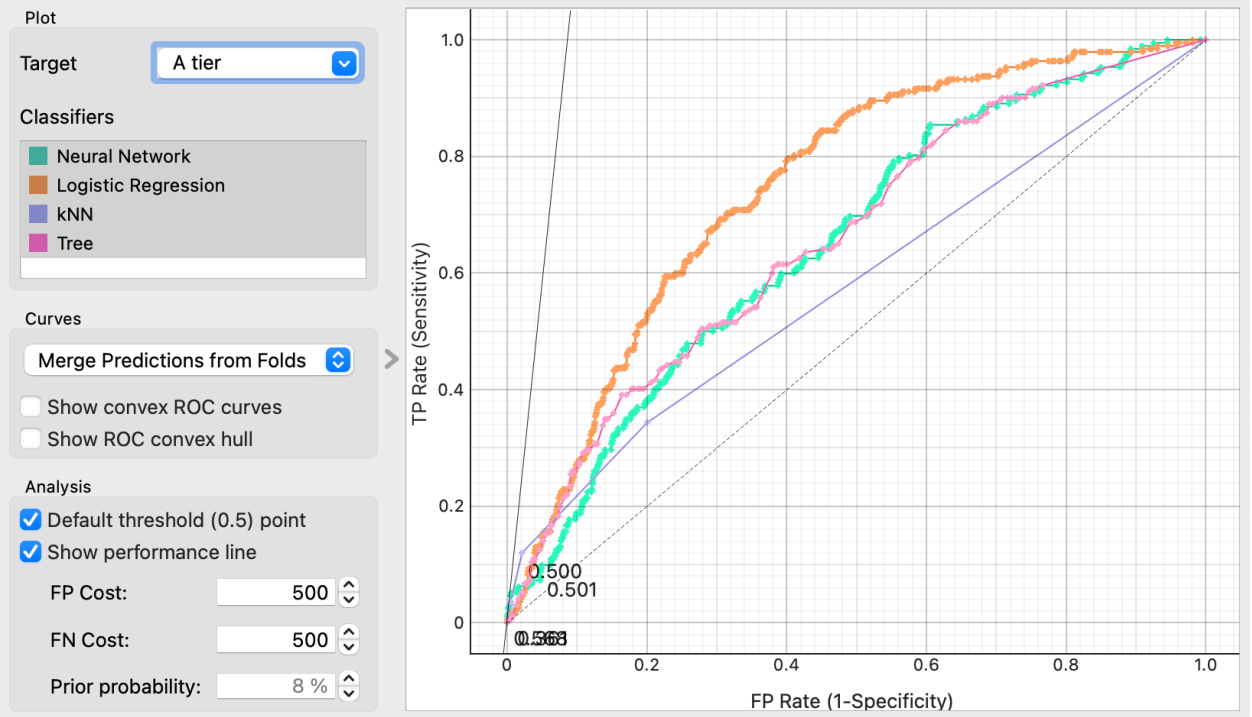


Рисунок 9 – Скріншот графіку ROC-кривих для класу 1 (A tier)

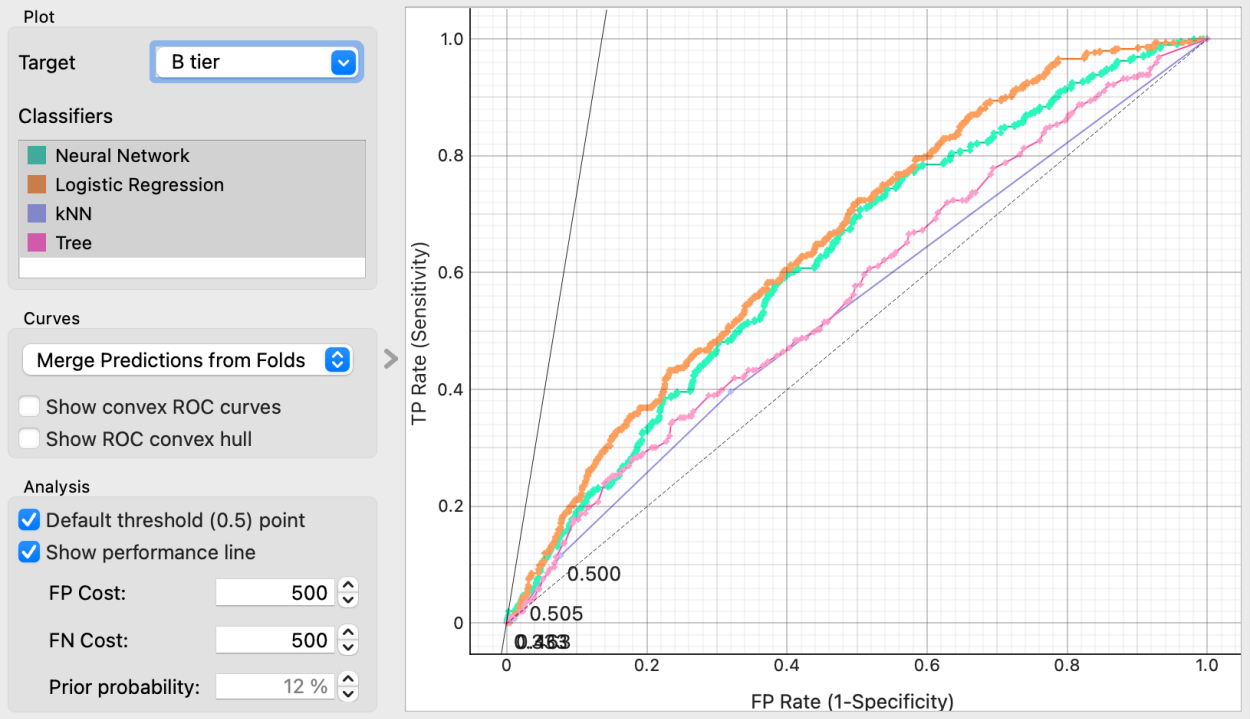


Рисунок 10 – Скріншот графіку ROC-кривих для класу 2 (B tier)

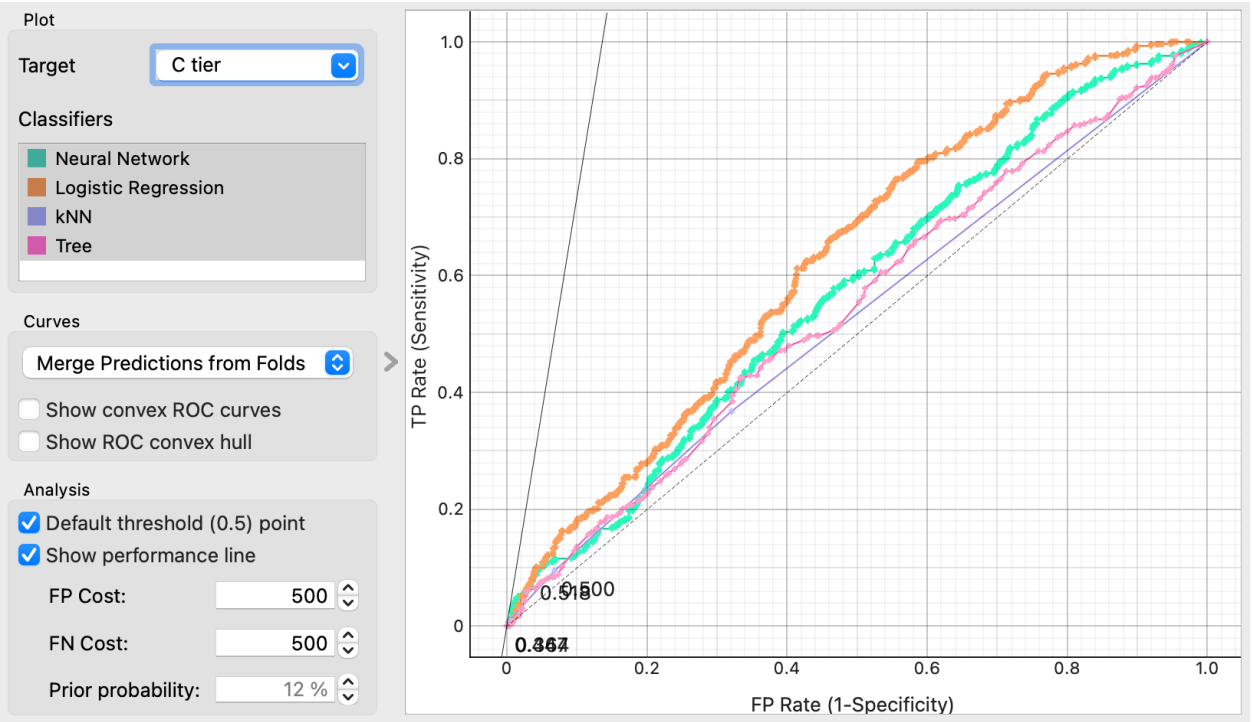


Рисунок 11 – Скріншот графіку ROC-кривих для класу 3 (C tier)

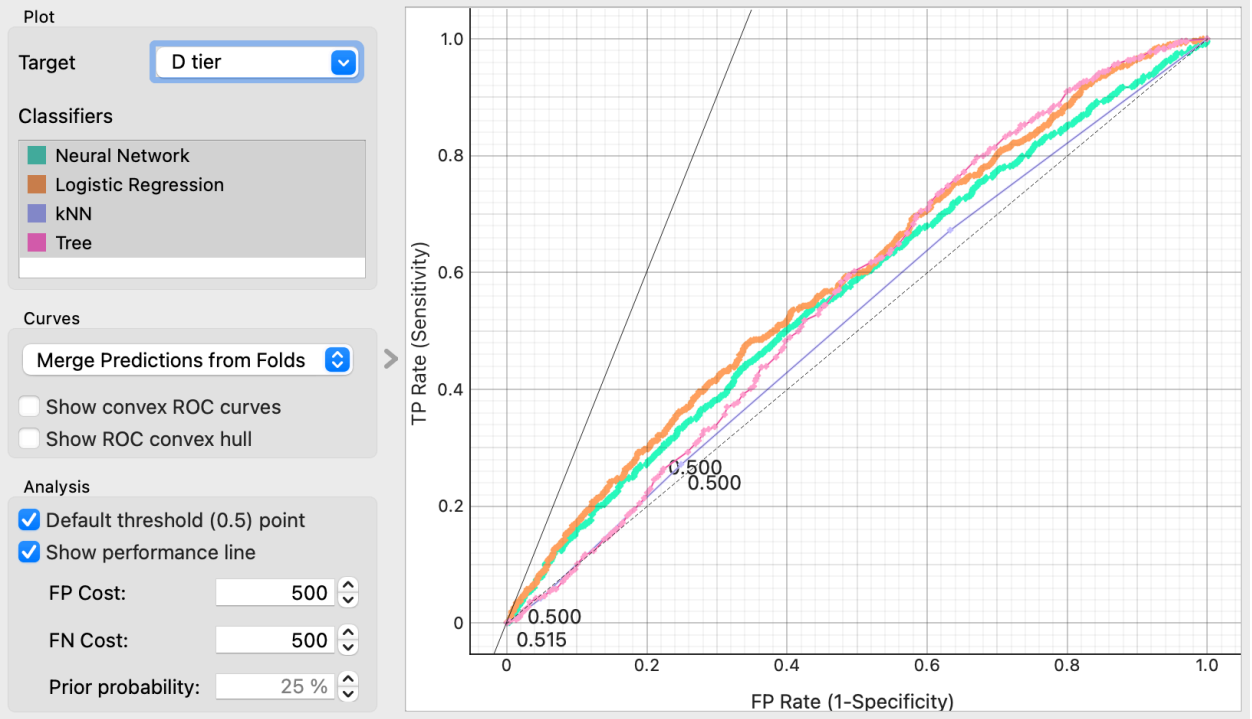


Рисунок 12 – Скріншот графіку ROC-кривих для класу 4 (D tier)

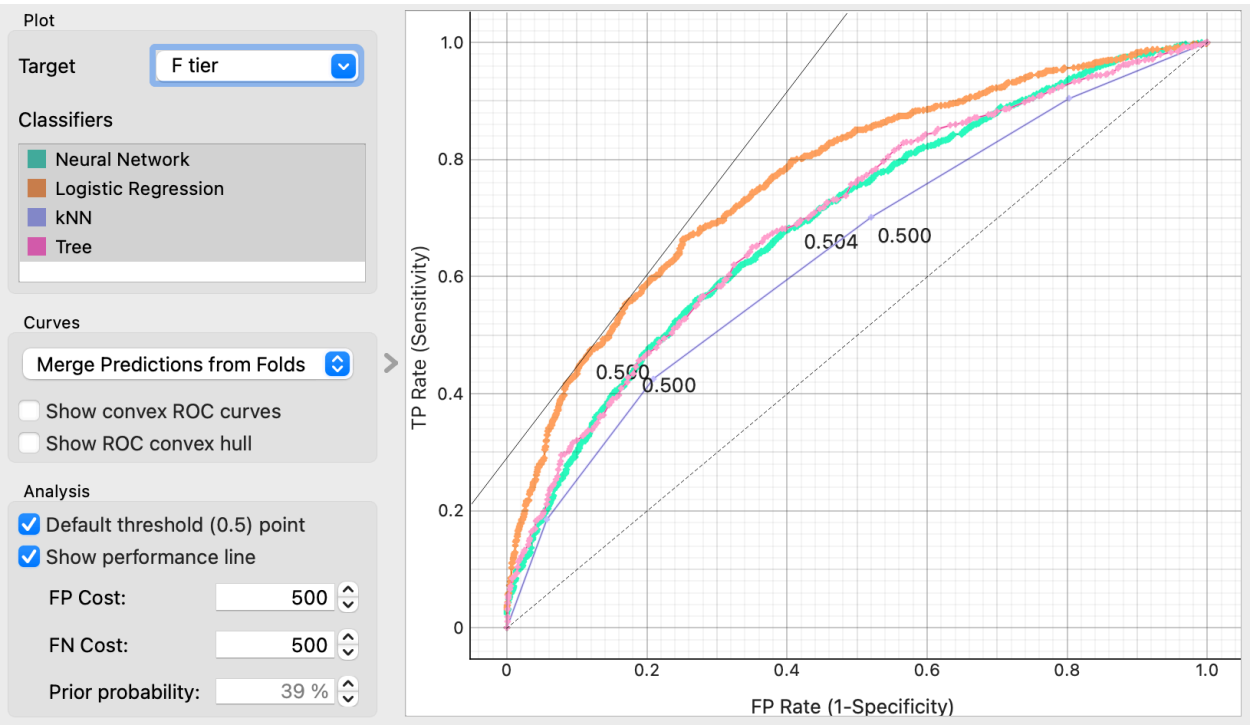


Рисунок 13 – Скріншот графіку ROC-кривих для класу 5 (F tier)

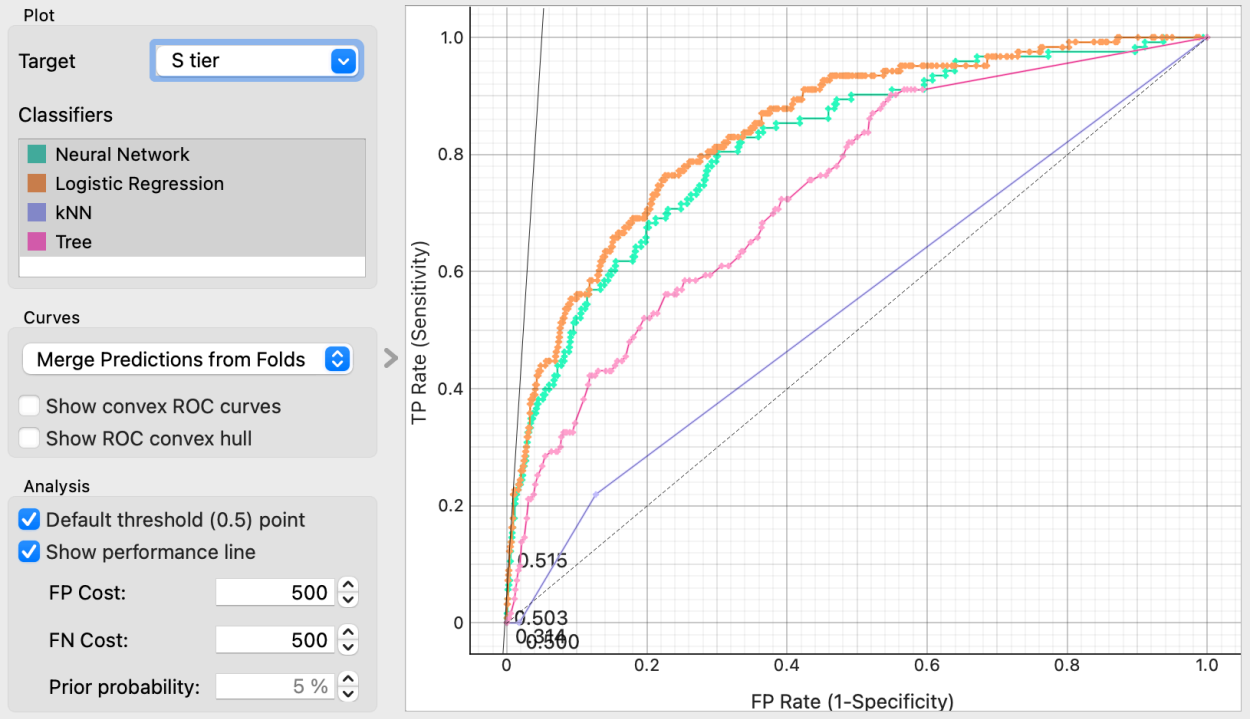


Рисунок 14 – Скріншот графіку ROC-кривих для класу 6 (S tier)

Графіки ROC-кривих показують, що класи 3 і 4 (C tier і D tier відповідно) мають погану якість розпізнавання – моделі часто плутають позитивні та негативні приклади, що видно з форми кривих. У той час як клас 6 (S tier) розпізнається найкраще, а інші класи мають середню якість.

Продивимось результати віджету Confusion Matrix для найкращої класифікаційної моделі (Linear Regression) при при параметрі Show = Number of Instances (рис. 15) та при Show = Proportion of Actual (рис. 16).

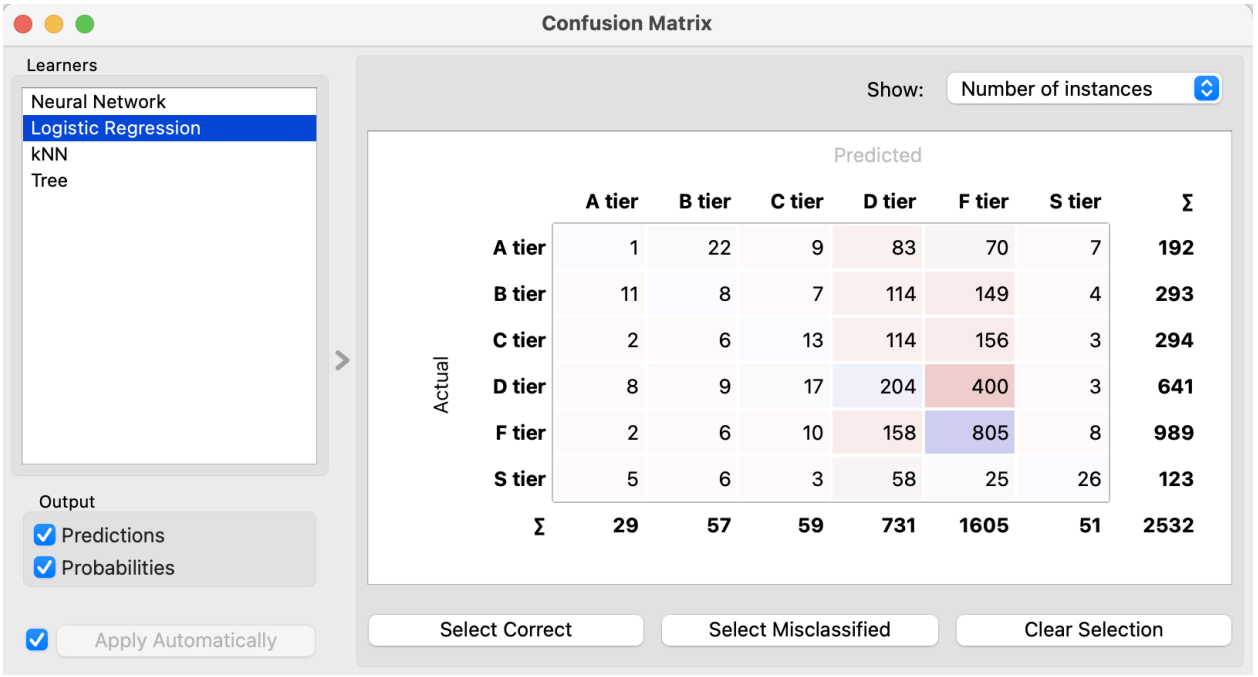


Рисунок 15 – Скріншот віджету Confusion Matrix для логістичної регресії при параметрі Show = Number of Instances

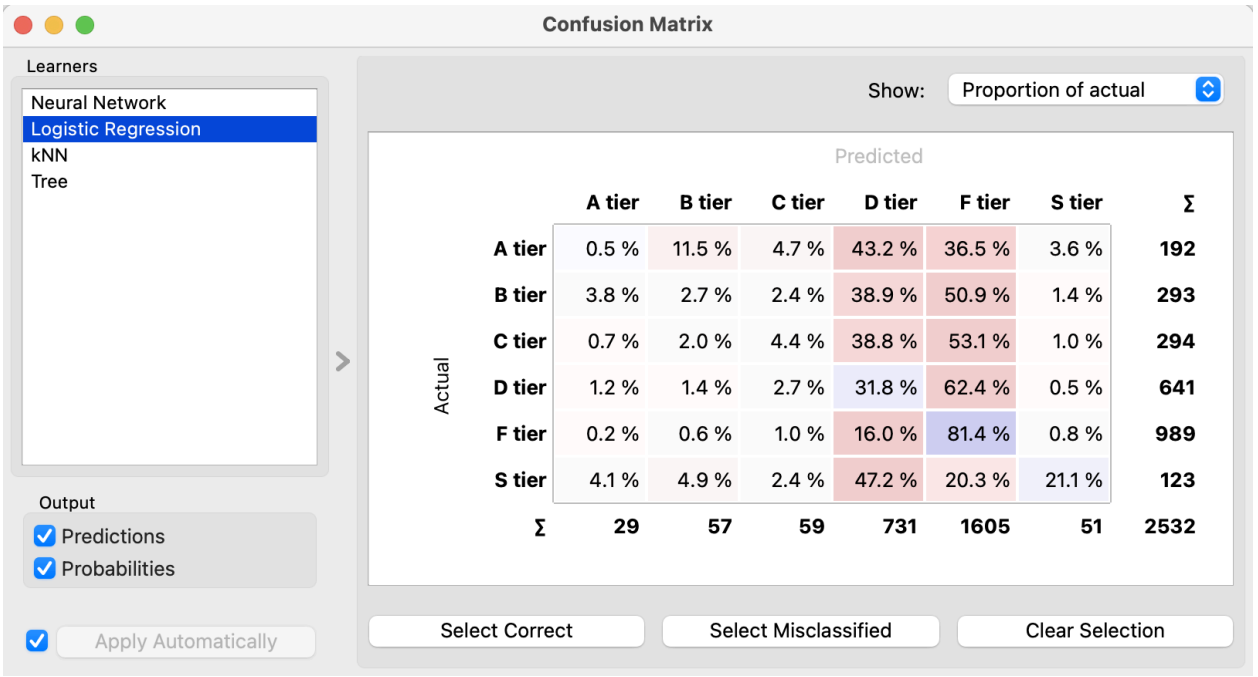


Рисунок 16 – Скріншот віджету Confusion Matrix для логістичної регресії при параметрі Show = Proportion of Actual

В результаті видно, що найкраща точність спостерігається для F tier, тоді як клас D tier часто плутається з класом F tier, а класи B tier та C tier

мають суттєве перекриття класом F tier. Отже, модель часто помилково прогнозує клас F tier, тому потребує покращення роздільності B tier та C tier.

Для інших моделей класифікації (Neural Network, kNN та Tree) збережемо лише скріншоти при параметрі Show = Proportion of Actual (рис. 17-19).

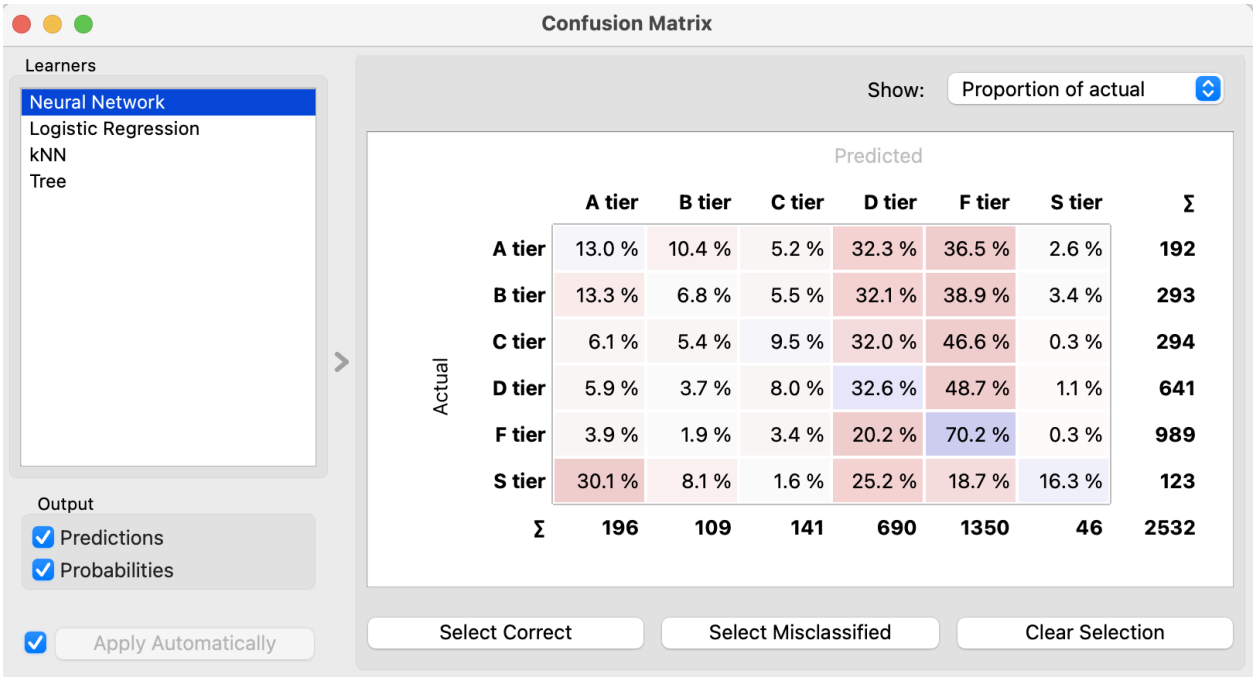


Рисунок 17 – Скріншот віджету Confusion Matrix для нейронної мережі при параметрі Show = Proportion of Actual

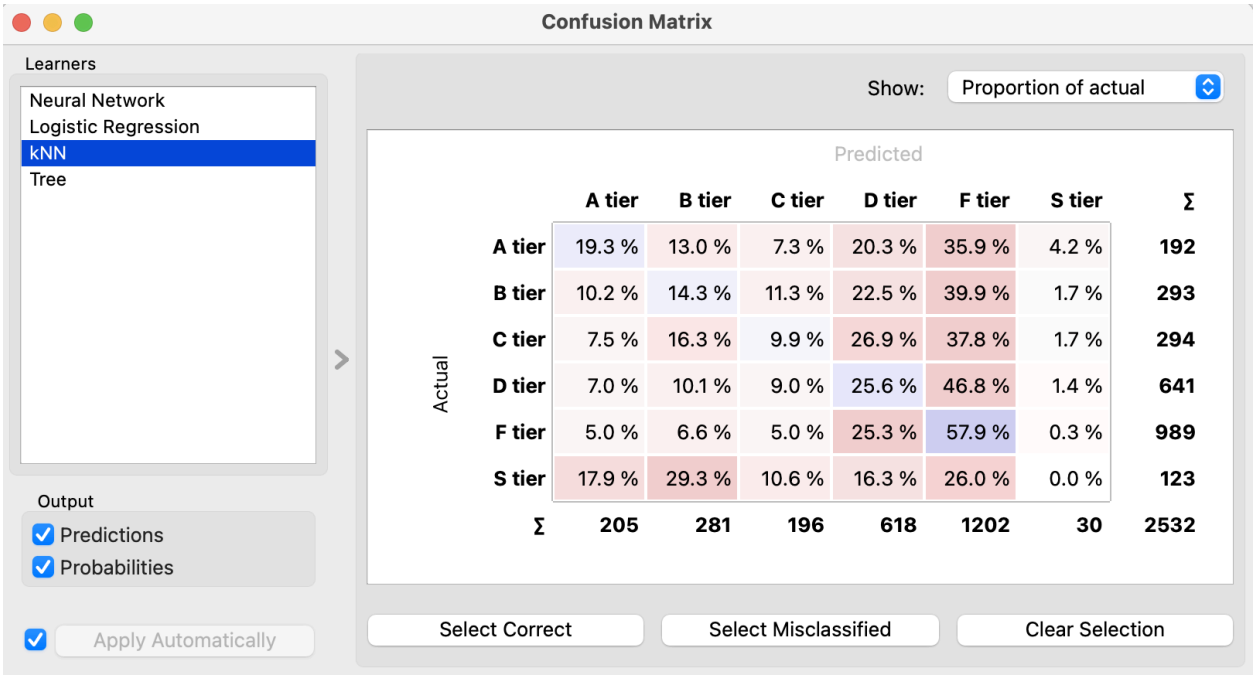


Рисунок 18 – Скріншот віджету Confusion Matrix для методу k найближчих сусідів при параметрі Show = Proportion of Actual

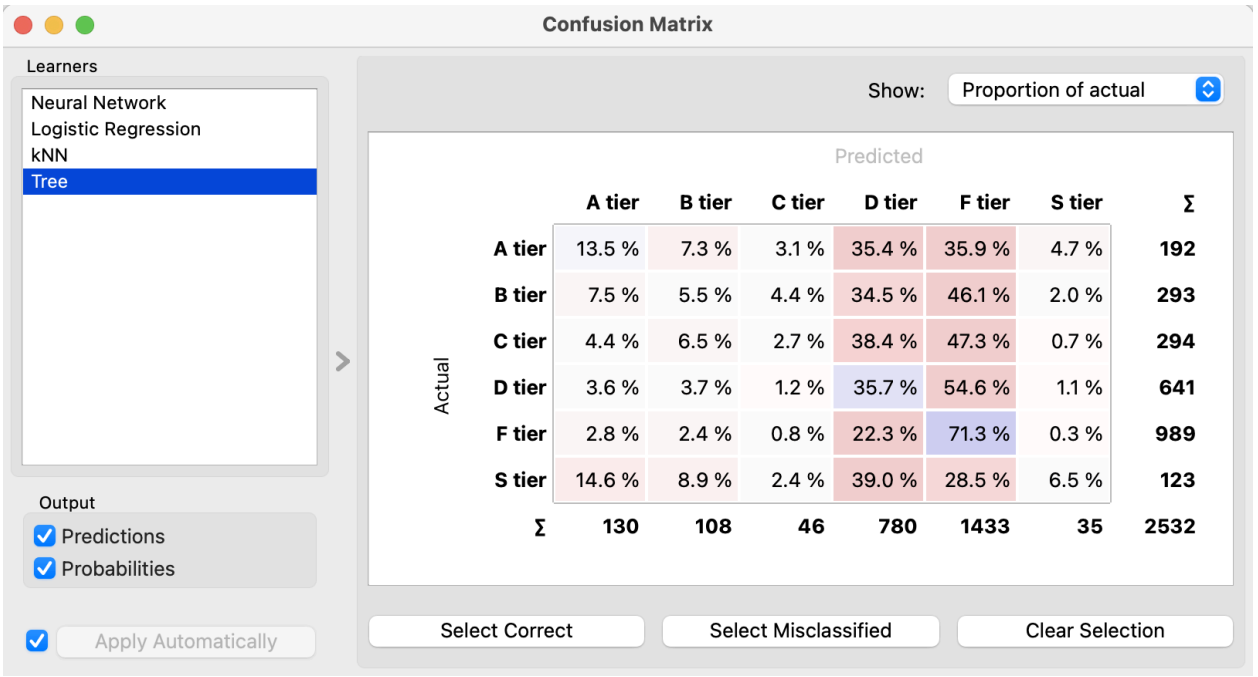


Рисунок 19 – Скріншот віджету Confusion Matrix для дерева рішень при параметрі Show = Proportion of Actual

Цікаво, що, хоча kNN за точністю є найгіршою моделлю для даної задачі, вона демонструє найкращу точність для класів A tier та B tier.

Висновки. В ході виконання розрахунково-графічної роботи були закріплені знання на тему класифікації об’єктів за ознаками. Було проведено моделювання задачі класифікації засобами пакета Orange.

Для того, щоб виконати класифікацію, була сформована вибірка та обрано чотири класифікатори, які використовувалися в лабораторній роботі №4. Підбираючи налаштування кожного класифікатора, була досягнута найвища можлива якість класифікації.

Результат тестування показав, що алгоритм логістичної регресії надав найкращий результат класифікації даних, а метод k-середніх – найгірший, але більш збалансований за класами. Це відбувається тому, що на відміну від логістичної регресії або дерева класів, які можуть підлаштовуватись під домінантний клас і фактично нехтувати малими категоріями, kNN дивиться на локальне оточення точок – його логіка не узагальнює дані жорстко, а лише повторює локальні патерни. В цілому, AUC був в межах від 62,5% до 69%. Такі відносно низькі результати були отримані через специфіку даних, з якими працювала модель. У задачі прогнозування бюджету фільмів за жанром, країною виробництва, студією та тривалістю отримати високі значення AUC практично неможливо, оскільки між цими ознаками й цільовою змінною немає прямих залежностей: один і той самий жанр може мати як надзвичайно дешеві, так і надзвичайно дорогі фільми, країна виробництва майже не впливає на витрати, а тривалість та студії дають лише слабкі кореляції. Логістична регресія змогла вловити ці мікропатерни і на основі їх суми досягти достатньо високого результату, у той час як інші моделі справлялися гірше, страждаючи то від перенавчання, то від надто слабких залежностей. Тому отриманий діапазон вважається нормальним та реалістичним результатом для подібного датасету.

Також, проглянувши графіки ROC-кривих та матриці помилок, ми побачили, що найточніші прогнози мають класи F tier та S tier. Це також цілком очікуваний результат, адже фільми з дуже високим та дуже низьким бюджетом мають сильніші кореляції. Наприклад, фільми жанрів фантастики,

бойовика та фентезі, особливо зняті відомими студіями, у більшості випадків мають великий бюджет, у той час як документальні фільми, особливо зняті невеликими країнами, такими як Білорусь чи Чеська Республіка, частіше мають малий бюджет. Класи C tier та D tier практично не прогнозуються, адже середній бюджет має надто великий діапазон використання, тому кореляція майже відсутня.